The Impact of Aggregated Leaks on Privacy

Graduate

Simon Kindhauser

Caspar Martens

Advisor
Prof. Dr. Mitra Purandare

Co-Examiner
Dr. Andreas Wespi, IBM Research - Zurich, Niederhasli, ZH

Subject Area
Security, Software, Miscellaneous

Introduction: "Das Internet ist für uns alle Neuland". (The internet is uncharted territory for all of us.) - Angela Merkel, former German Chancellor

This statement is arguably still true. We still have no reference of what happens after a lifetime of usage. As we spend more time on the internet it gets increasingly likely that a used service loses control over the data it collected. Over the decades, we add more and more data to our digital footprint. The likelihood of our private information being exposed increases, leading to potential consequences that are difficult to predict. This raises the question of whether the surging availability of personally identifiable information will increase the possibility of linking such information together and how this could be prevented.

Approach / Technology: To answer these questions we developed two complimentary approaches for analyzing a synthetic dataset based on data leaks in the real world. For this we use metadata provided by "Have I been Pwned" describing over 650 distinct leaks containing a total of 12 billion records. We further enrich this information with statistical knowledge from other sources regarding the distribution of attributes.
Both approaches analyze the data using the linkage algorithms depicted in figure 1 as well as various different mitigation strategies. They produce statistical information regarding the accuracy of the algorithms as well as the effectiveness of the mitigation strategies using the Average Discovery Ratio as illustrated in figure 3.
The first approach generates a subset of the entire dataset with only US citizens based on a Snakemake pipeline and Python library. This dataset is then loaded into Neo4j for analysis where a user-defined procedure was created, implementing the first and second order algorithms. With this approach we are able to verify our algorithms and record the true positive, true negative and false negative values as shown in figure 2.
The second approach simulates all records of a person -- their entire attack surface -- within the synthetic dataset in real time. The advantages being a generally lower memory footprint whilst providing competitive performance. This allows for the implementation of all three linkage algorithms. The main limitation is that false positives are not possible which results in an incapacity to measure the accuracy of the algorithms.

Result: We show that our linkage algorithms can link 90% of leaked records, effectively creating clusters belonging to a specific person. We evaluated different mitigation strategies regarding their levels of protection against the presented attacks as can be seen in figure 3. Using unique email addresses is the most effective, single mitigation strategy against first order linkage. However, against more sophisticated attacks we recommend a combination of unique email addresses and passwords as well as User-Initiated Differential Privacy. Using this combined mitigation strategy can decrease the linkage effectiveness by up to 50% while being relatively easy to implement thanks to existing and well-established tooling such as password managers.

Figure 1: Linkage Algorithms
Own presentment

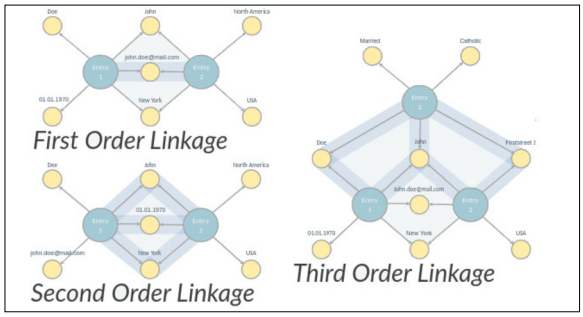

Figure 2: Confusion Matrix of Linkage Algorithms
Own presentment

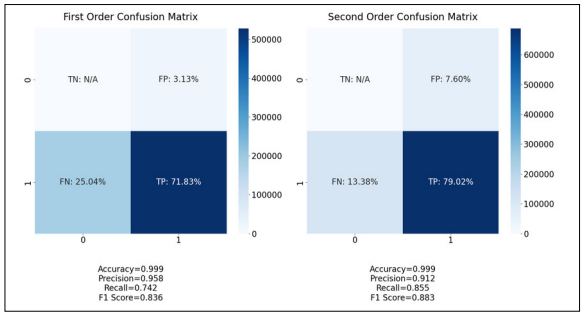

Figure 3: Effectiveness of Mitigation Strategies
Own presentment